DOCUMENT RESUME

ED 387 501 TM 023 637

AUTHOR	Kino, Mary M.; And Others
TITLE	Differential Objective Function.
PUB DATE	Apr 95
NOTE	34p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Ability; Analysis of Variance; Demography; Difficulty Level; Estimation (Mathematics); Ethnicity; Grade 8; *Identification; *Item Bias; Item Response Theory; Junior High Schools; *Mathematics Tests; Profiles; *Scores; Sex Differences; *Test Items
IDENTIFIERS	*Connecticut Mastery Testing Program; *Differential Objective Function

ABSTRACT

Item response theory (IRT) has been used extensively to study differential item functioning (dif) and to identify potentially biased items. The use of IRT for diagnostic purposes is less prevalent and has received comparatively less attention. This study addressed differential objective function (dof) to identify potentially biased content units. IRT was used to estimate person abilities and item difficulties, which were used to compute residual objective scores. Residual objective scores were analyzed with analysis of variance using the independent variables gender and ethnicity. Data were from mathematics subtests from the 1992 Connecticut Mastery Test census administration of eighth graders and its database of approximately 32,000 Connecticut eighth graders. The examples illustrate how dof outcomes can be used to identify potentially biased content units, to provide diagnostic information at the content level, and to construct profiles of content-based performance for different demographic subgroups. Ten figures and two tables present analysis results. Two appendixes present dif statistics by demographic subgroup and item-level statistics for dof objectives in four tables. (Contains 11 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 387 501

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
MARY M. KINO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Differential Objective Function

Mary M. Kino
Huixing Tang
David Swift

The Psychological Corporation

Paper presented at the 1995 Annual Meeting of
the National Council on Measurement in Education

San Francisco, California

TM023437

ACKNOWLEDGEMENT: The authors gratefully acknowledge the Connecticut State Department of Education for access to the 1992 Connecticut Mastery Test data used in this study.

BEST COPY AVAILABLE

ABSTRACT

Item response theory (IRT) has been used extensively to study differential item functioning (*dif*) and to identify potentially biased items. The use of IRT for diagnostic purposes is less prevalent and has received comparatively less attention.

This study addresses differential objective function (*dof*) to identify potentially biased content units. IRT was used to estimate person abilities and item difficulties, which were used to compute residual objective scores. Residual objective scores were analyzed with analysis of variance using the independent variables gender and ethnicity. The examples illustrate how *dof* outcomes can be used to identify potentially biased content units, to provide diagnostic information at the content level, and to construct profiles of content-based performance for different demographic subgroups.

RATIONALE

Applications of item response theoretic (IRT) methods have enhanced the process of test development and test construction (Hambleton, 1989), evolutionized computer-adaptive testing technology, and facilitated test equating procedures. The item function is defined by simultaneously estimating person and item parameters, and is expected to be comparable between matched-ability groups that differ on characteristics independent of ability. IRT-based methods present significant contributions to the investigation of differential item functioning (*dif*) and potential item bias. (See, for example, reviews by Ironson, 1983, and Shepard, Camilli, & Averill, 1981.) IRT-based methods have additionally been used to test the sufficiency of model-data fit and its relationship to potential item bias (Linn & Harnisch, 1981; Wright, Mead, & Draba, 1976).

Despite the widespread use of IRT in technical areas of test development, its application for curricular diagnosis and content analysis is less prevalent. Popular *dif* methods use discrete items as the unit of analysis. Although those types of analyses serve multiple purposes, interpretations about the test content are not inherently tied to those methods. Identifying potential bias through differential function at any level -- item, objective, or other content-based units -- is ultimately a function of a substantive content review. Traditional methods typically address only one factor at a time and ignore interaction effects from multiple *dif* factors like gender and ethnicity. Ignoring *dif* interaction effects can result in misleading interpretations about *dif* main effects.

Tang (1994) proposed an IRT-ANOVA method which addresses simultaneous *dif* analysis for multiple levels and multiple factors. IRT is used to estimate person ability and item difficulty parameters. Residual scores, free from the effects of person ability and item difficulty, are computed. Differences in residual scores between different demographic groups, defined by different levels of *dif* factors, are then tested with analysis of variance. Any significant differences in the demographic groups' mean residual scores may be an indication of potential bias.

The current study extends the IRT-ANOVA method's unit of analysis from the discrete item level to the content-based unit. The empirical analysis of content-based units contextualizes the statistical significance of discrete *dif* items. This content-based extension presents several advantages over traditional *dif* methods:

i. the analyses are performed on content-based units;
ii. the method can simultaneously address multiple levels and multiple factors;
iii. interaction effects can be studied while controlling for confounding variables;
iv. the outcomes lend themselves readily to content-based interpretations; and
v. content-based interpretations are more amenable to diagnostic applications.

The content units may be defined by curricular objectives, content domains, or other substantive units which are used to define test content. The content unit in this study is the curricular objective and its analysis is referred as differential objective function.

Differential objective function (*dof*) occurs when objectives function differently for particular subgroups of examinees irrespective of underlying ability. The presence of *dof* may be attributed to differences in opportunity to learn (Lehman, 1986; Muthén, Kao, & Burstein, 1991), in instructional bias (Linn & Harnisch, 1981), or in other curricular factors. Lower levels of performance may be attributed to differences in instructional delivery and in opportunity to learn. Given the tenability of model assumptions, differences in item performance between matched-ability groups are indicative of *dif*. *Dof* is more likely than item-level *dif* to yield content-based explanations about the observed differences between matched-ability groups. Outcomes at the objective level can provide collateral information which otherwise remains untapped from outcomes of discrete items alone.

The results from this study illustrate how *dof* can inform interpretations of item analysis: It augments *dif* data, contextualizes the significance of item statistics, and provides diagnostic information at the objective level.

METHOD

SAMPLE

The current study is a secondary analysis of mathematics subtest data from the 1992 Connecticut Mastery Test census administration of eighth-graders in Connecticut public schools. The mathematics subtest consisted of 144 dichotomously-scored multiple choice items. These items measured mathematics performance on 36 curricular objectives, each comprised of four items.

Two dichotomized student background variables -- gender (Female/Male) and ethnicity (Black/White) -- were the *dof* factors and formed the sampling strata for the study. From the database of approximately 32,000 Connecticut eighth-graders, item responses and demographic data from 400 examinees were randomly sampled from each (gender × ethnicity) demographic stratum to yield a total sample size of 1600.

LIMITATIONS OF THE STUDY

This study is a secondary analysis of an existing data set which does not include information about methods of instructional delivery, opportunity to learn, or instructional bias. The current analyses exclude attempts to validate the interpretation of *dof* as a function of any of these factors. The methods described in this study are reported as part of developmental work in an area which warrants further consideration and continued research.

PROCEDURE

At the objective level, expected performance was modeled as a function of examinee ability and difficulty of the objective. Residual objective scores are a function of item scores adjusted for person ability and item difficulty, and reflect the difference between the expected and observed objective scores. They are expected to be random with a mean of 0. A positive (or negative) residual implies that an examinee's score is higher (or lower) than expected. Consistently high (or low) residuals for a subgroup imply that the objective favors (or disfavors) the subgroup.

The procedure applies a one-parameter logistic IRT model to dichotomously-scored items. Item responses are assumed essentially unidimensional and locally independent within and across objectives. The initial steps at the individual examinee level for person n (1, . . ., N); item i (1, . . ., I_j) nested within objective j (1, . . ., J) are:

Step 1. Calibrate the data for the intact group. Obtain estimates of person ability (B_n) and item difficulty (D_i).

Step 2. Use the estimates obtained in Step 1 to compute person n's expected item i score, $E_{ni} = \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)}$. The observed item i score for person n is X_{ni}. For dichotomously-scored items, $X_{ni} = 1$ if correct, 0 otherwise.

Step 3. Compute person n's expected objective score by adding the expected item scores nested within objective j, $E_{nj} = \sum_{i=1}^{I_j} E_{ni(j)}$. The observed objective score for person n is the sum of the item scores nested within objective j, $X_{nj} = \sum_{i=1}^{I_j} X_{ni(j)}$.

Step 4. Compute person n's residual objective score, $R_{nj} = X_{nj} - E_{nj}$. R_{nj} is the difference between the observed and expected objective scores, and it reflects the magnitude of *dof* for person n on objective j.

Step 5. Apply analysis of variance on the R_{nj}'s as the dependent variable and *dof* factors (gender and ethnicity) as the independent variables.

The generalized linear model is:

$$\underline{\mathbf{R}}_j = \underline{\mathbf{X}}\underline{\beta}_j + \underline{\varepsilon}_j$$

where $\underline{\mathbf{R}}_j$ = [$N \times 1$] vector of N person residuals, R_{nj}, for objective j;

$\underline{\mathbf{X}}$ = [$N \times G$] "design" matrix of N persons' values on each independent *dof* variable in the model;

$\underline{\beta}_j$ = [$G \times 1$] vector of regression coefficients for objective j; and

$\underline{\varepsilon}_j$ = [$N \times 1$] vector of N persons' error terms for objective j.

In this study, $\underline{\beta}_j$ takes on the form $[\beta_0\ \beta_{ETHNIC}\ \beta_{GENDER}\ \beta_{ETHNIC \times GENDER}]'$.

Step 6. Compute the residual mean objective scores for mutually exclusive demographic subgroups, defined by the levels of the *dof* factors.

The residual mean objective score reflects the magnitude of *dof* for the demographic subgroup. For example, a residual mean objective score of 0.15 for a subgroup indicates that the group as a whole performed better than expected by 0.15 objective score points, given the group's ability level and the difficulty of the objective.

RESULTS

Residual objective scores were modeled via general linear models, with *dof* factors gender and ethnicity as independent variables. *Dof* main effects and two-way interactions were tested for significance using the univariate F-ratio as the *dof* test statistic. The magnitude of residual mean difference was used as an additional criterion for significant *dof*. Appendix A presents residual mean objective scores and magnitudes of residual mean difference by main effects gender and ethnicity, two-way (gender × ethnicity) interactions, their univariate F-ratios, and corresponding *p*-levels of significance. Univariate F-ratios were computed separately for each objective.

Significant *dof* was detected on 10 of the 36 objectives for main effects and 2-way interactions at the $\alpha = 0.01$ level. For main effect *dof*, an additional criterion of difference

in group residual mean objective scores greater than or equal to 0.15 was applied. These results are summarized in Table 1.

Table 1
Summary of Significant *Dof*

Dof Effect	Number of Objectives
Main Effect Ethnicity*	4
Main Effect Gender*	7
2-way Interaction Ethnicity × Gender	2
Non-significant *dof*	26

*3 common objectives, significant *dof* main effects for ethnicity and gender

Although significant *dof* was detected for eight unique objectives at the main effects level, these outcomes should be interpreted in light of at least two considerations:

(a) The statistical significance of main effects could be attributed to increased power and larger sample size (n = 400 examinee responses for each 2-way interaction effect, compared to n = 800 examinee responses for each main effect).

(b) Error rates of significance tests increase with repeated significance tests performed on the same sample.

Subsequent discussion of the results and examples of *dof* are limited to two-way *dof* interactions.

Dof information and item-level *dif* data can enhance content-based interpretations. Three objectives -- Objective 3 with non-significant *dof*, Objectives 10 and 14 with statistically significant *dof* -- are highlighted to show how *dof* interactions can be interpreted. Two-way *dof* plots for the three objectives appear as Figures 1-3. Neither of the two-way (ethnic × gender) plots for Objective 3 [Figures 1(a) and 1(b)] shows a significant interaction effect at the objective level. Inspection of the objective level data reveals no apparent "gender gap" or "ethnicity gap."

DOF Plots

Figure 1(a) Objective 3, 'Gender Gap'

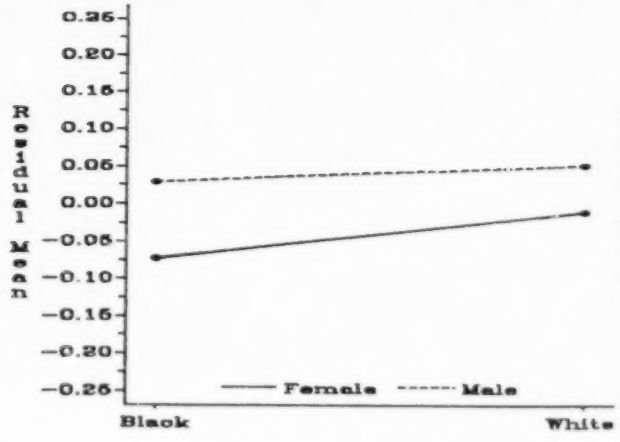


Figure 1(b) Objective 3, 'Ethnicity Gap'

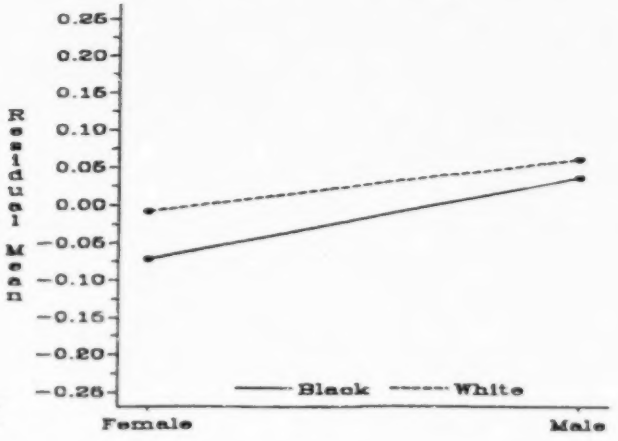


Figure 2(a) Objective 10, 'Gender Gap'

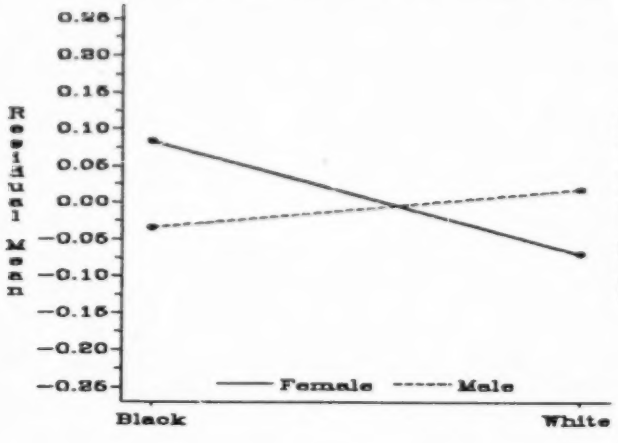


Figure 2(b) Objective 10, 'Ethnicity Gap'

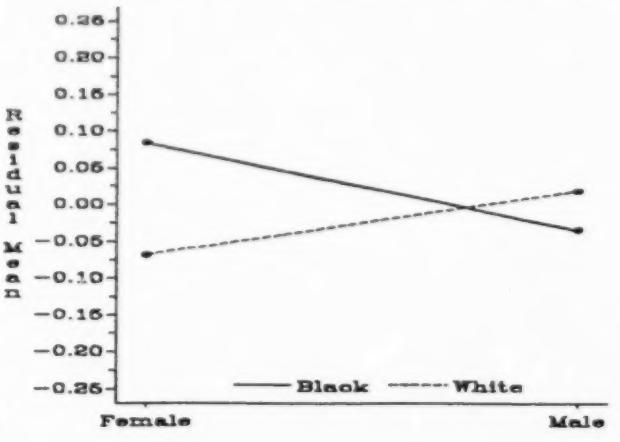


Figure 3(a) Objective 14, 'Gender Gap'

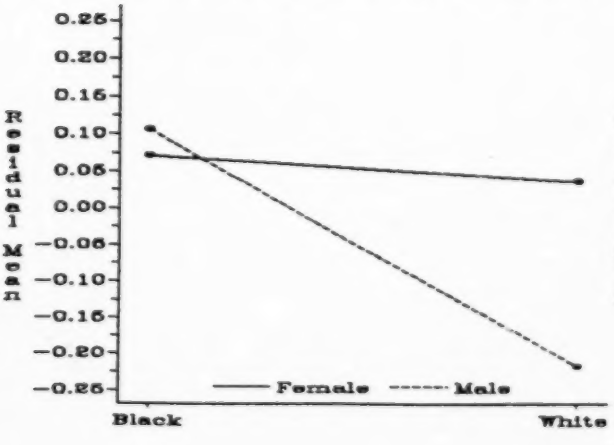
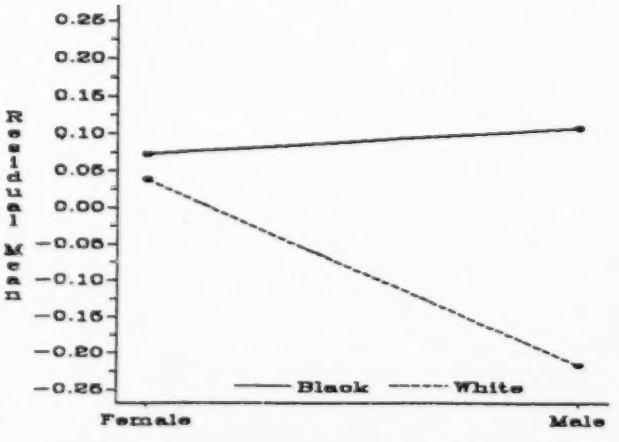


Figure 3(b) Objective 14, 'Ethnicity Gap'



Objectives 10 and 14 in Figures 2(a), 2(b), 3(a), and 3(b) illustrate two-way interactions. The item-level *dif* data for these objectives are presented in Appendix B. Objectives 10 and 14 were flagged with statistically significant (ethnicity × gender) *dof* interaction and appear to be of substantive significance. These *dof* interactions appear in Figures 2(a)-3(b).

The magnitude of the two-way *dof* interaction is operationalized by the difference between group differences. For Objective 10, that magnitude was 0.02 for the "gender gap" and 0.10 for the "ethnic gap." According to these methods, Objective 10's group-by-objective interaction is more pronounced for different ethnic groups of the same gender. Although Objective 10's two-way plots reveal interaction effects, the magnitudes of the interaction do not appear to be significant.

For Objective 14, the magnitude of the two-way *dof* interaction was 0.22 with "gender gap" interaction between Whites and Blacks [(Black Males - Black Females) vs. (White Males - White Females)], and 0.30 with "ethnic gap" interaction between Males and Females [(Black Males - White Males) vs. (Black Females - White Females)]. The difference in residual mean objective scores between White Males and White Females was greater than between Black Males and Black Females: The "gender gap" was more pronounced for Whites than for Blacks. The difference in residual mean objective scores between Black Males and White Males was greater than the difference between Black Females and White Females. The "ethnicity gap" was more pronounced among Males than among Females, and more distinct than the "gender gap."

To interpret the *dof* outcomes relative to the items that comprise an objective, two-way plots of item-level data are presented for each of Objectives 3, 10, and 14 in Figures 4-6. As shown in Figures 4(a)-4(d), none of the items (#105-108) associated with Objective 3 (round whole numbers) revealed a significant (ethnic × gender) interaction effect. For this objective, non-significant item-level *dif* was consistent with non-significant objective-level *dof*.

Item Level DIF Plots, Objective 3

Figure 4(a), Item 105

P=0.209

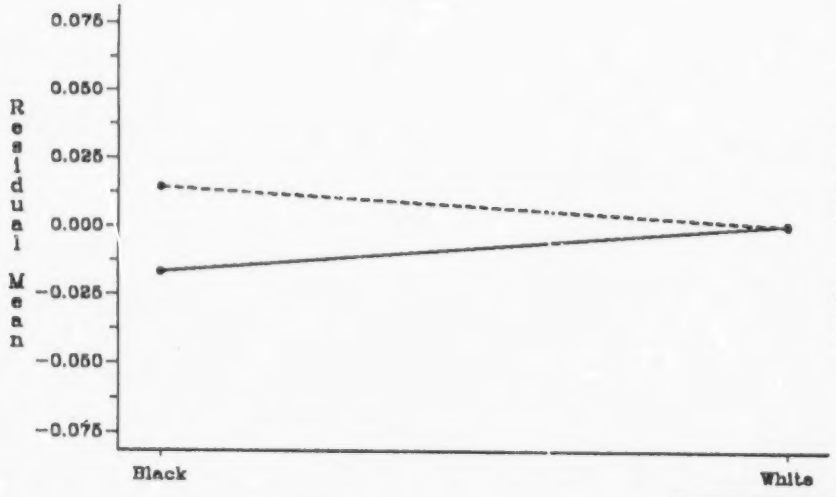
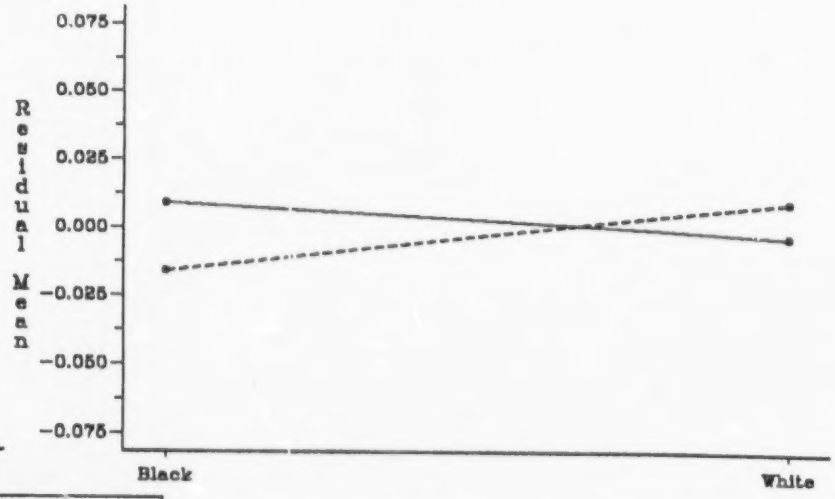


Figure 4(b), Item 106

P=0.161



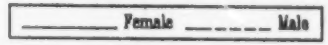


Figure 4(c), Item 107

P=0.53

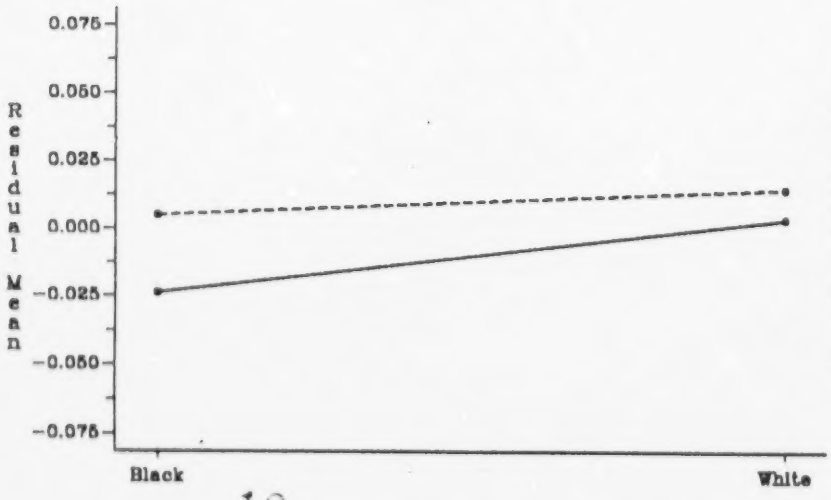
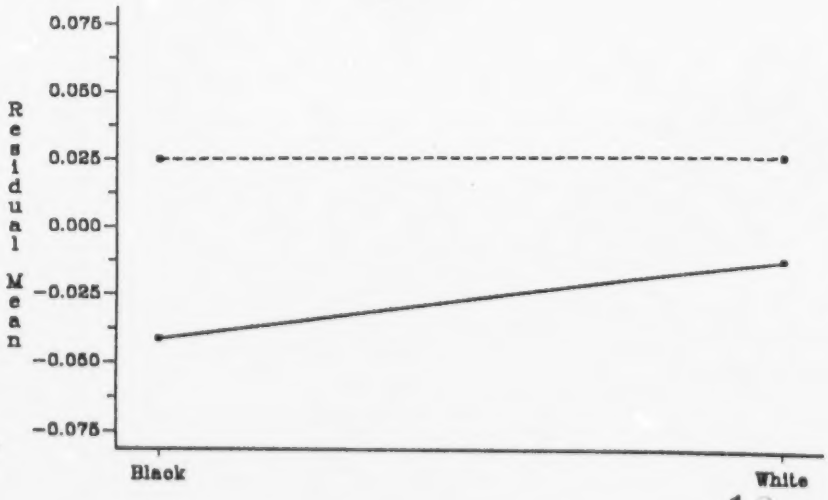


Figure 4(d), Item 108

P=0.471



Item Level DIF Plots, Objective 10

Figure 5(a), Item 113

P=0.056

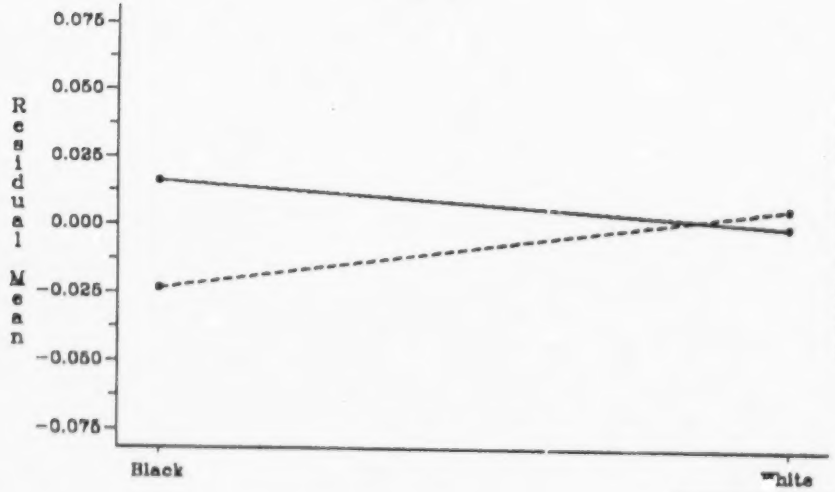
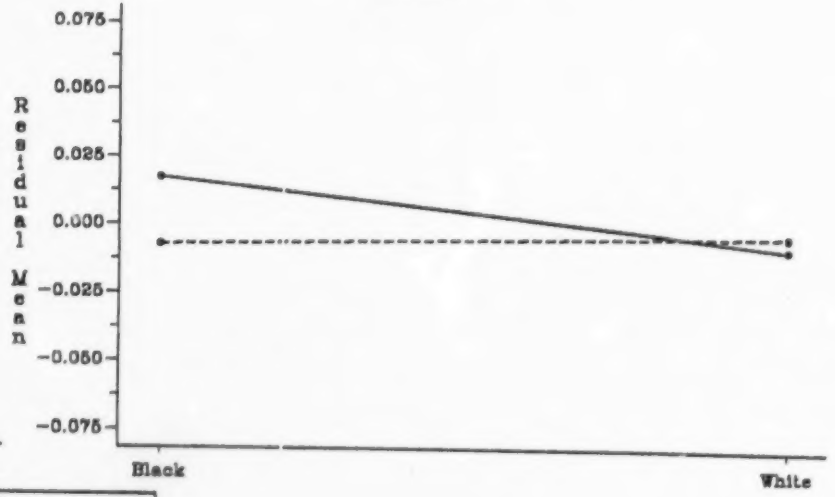


Figure 5(b), Item 114

P=0.364



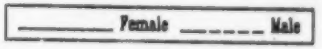


Figure 5(c), Item 115

P=0.379

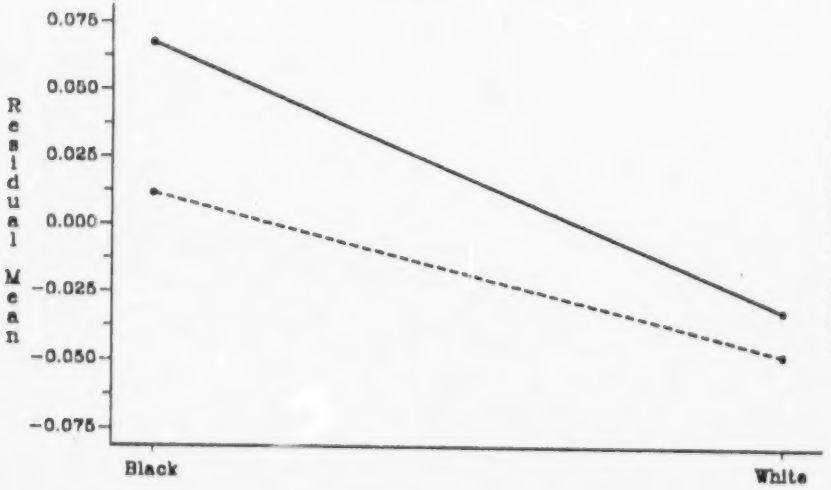
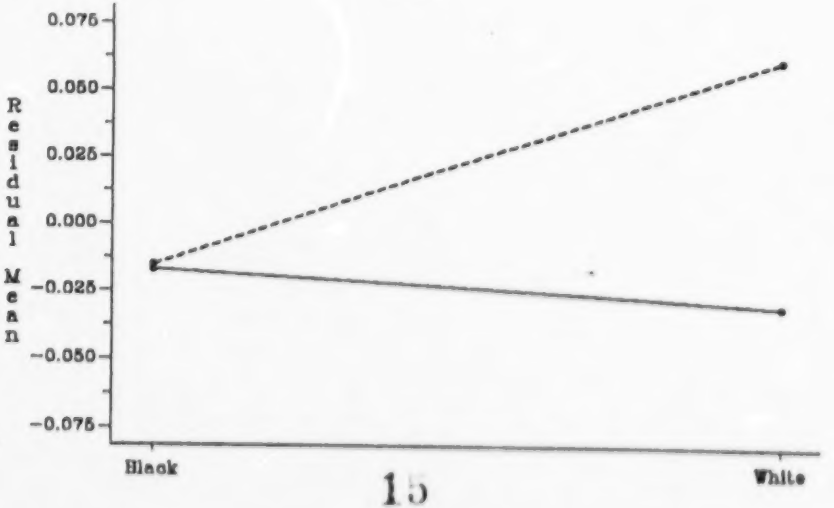


Figure 5(d), Item 116

P=0.031



Item Level DIF Plots, Objective 14

Figure 6(a), Item 65

P=0.028

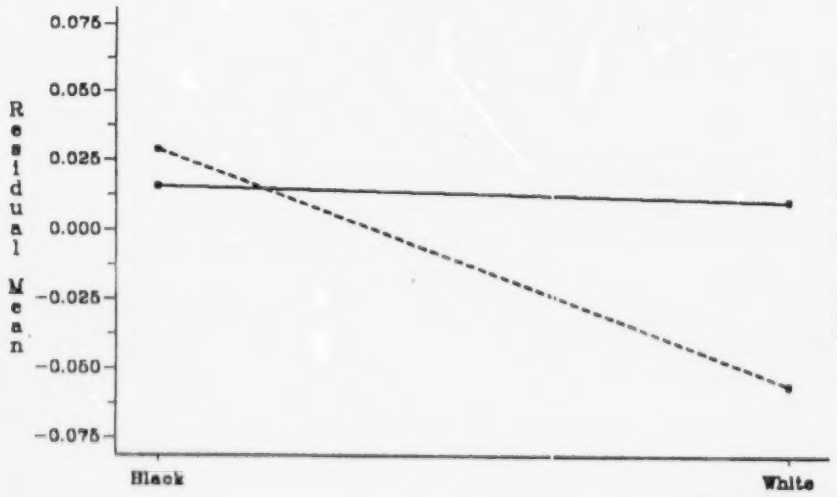
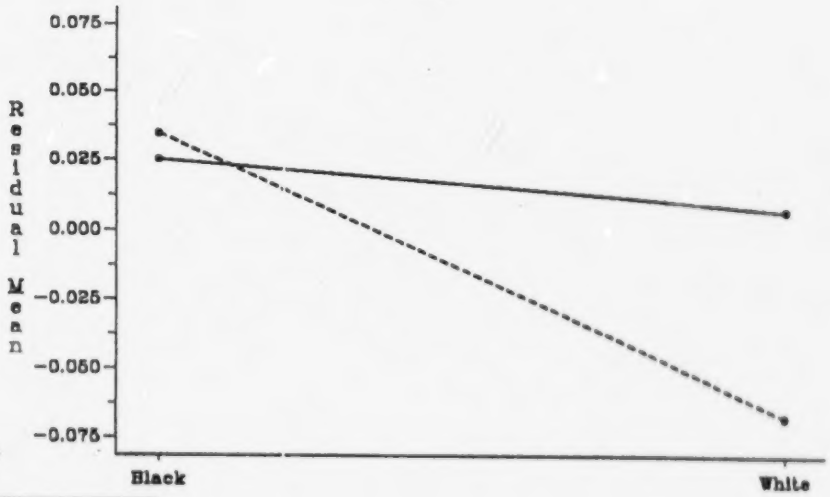


Figure 6(b), Item 66

P=0.013



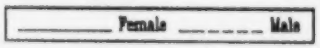


Figure 6(c), Item 67

P=0.069

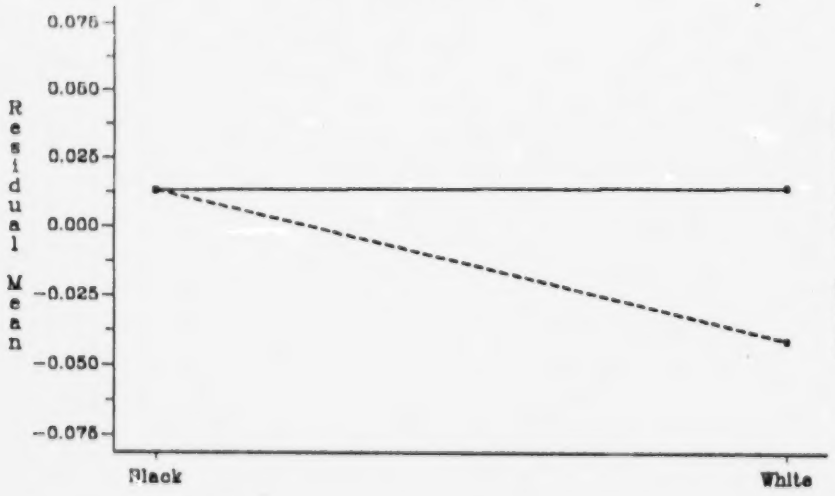
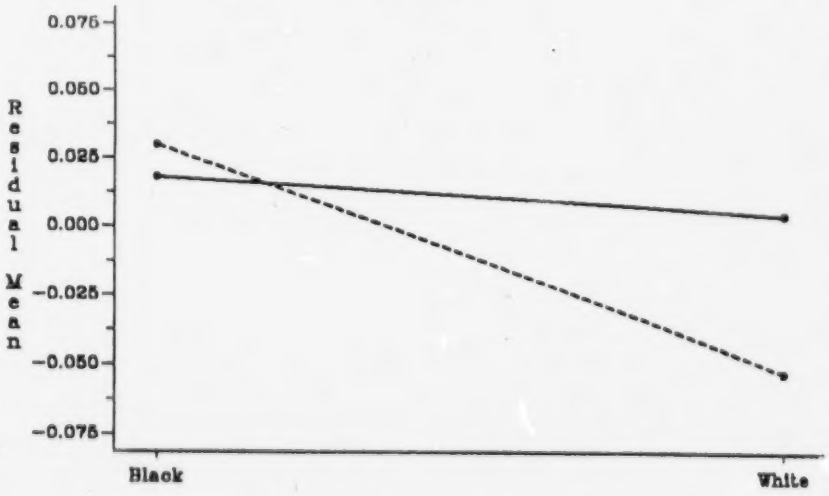


Figure 6(d), Item 68

P=0.041



Two-way interaction plots of the four items associated with Objective 10 (identify ratios and fractions from pictures) appear as Figures 5(a)-5(d). These plots show different patterns of interaction for the four demographic subgroups. Three of the four items (#113-115) show statistically non-significant interaction effects for the four demographic subgroups and appear to favor Black Females. Item #116 appears to favor White Males, while neither favoring nor disfavoring White Females, Black Females, or Black Males. The cumulative interaction effect of items #113-115, in addition to the interaction effect of item #116, may have resulted in the statistically significant *dof* interaction.

The item-level *dif* data associated with Objective 14 (add/subtract decimals to numbers of the form .XX), also flagged for a significant (ethnicity × gender) *dof* interaction, appear in Figures 6(a)-6(d) and show consistent interaction patterns between the four demographic subgroups. All four items (#65-68) consistently disfavored White Males and neither favored nor disfavored White Females, Black Females, or Black Males. The item-level and objective-level data are consistent. For this objective, and as measured by items #65-68, substantive content-based factors appear to differentiate the performance of White Males from other demographic subgroups.

GROUP PERFORMANCE PROFILES

Group performance profiles are presented in Figures 7-10. Each of the 36 objectives in this study was categorized into one of four content domains:

- conceptual understanding, Objectives 1 ~ 11;
- computational skills, Objectives 12 ~ 21;
- problem solving & application, Objectives 22 ~ 31; and
- measurement, Objectives 32 ~ 36,

partitioned by the vertical dotted lines in each of Figures 7-10.

Figure 7. WHITE MALE RESIDUAL MEANS

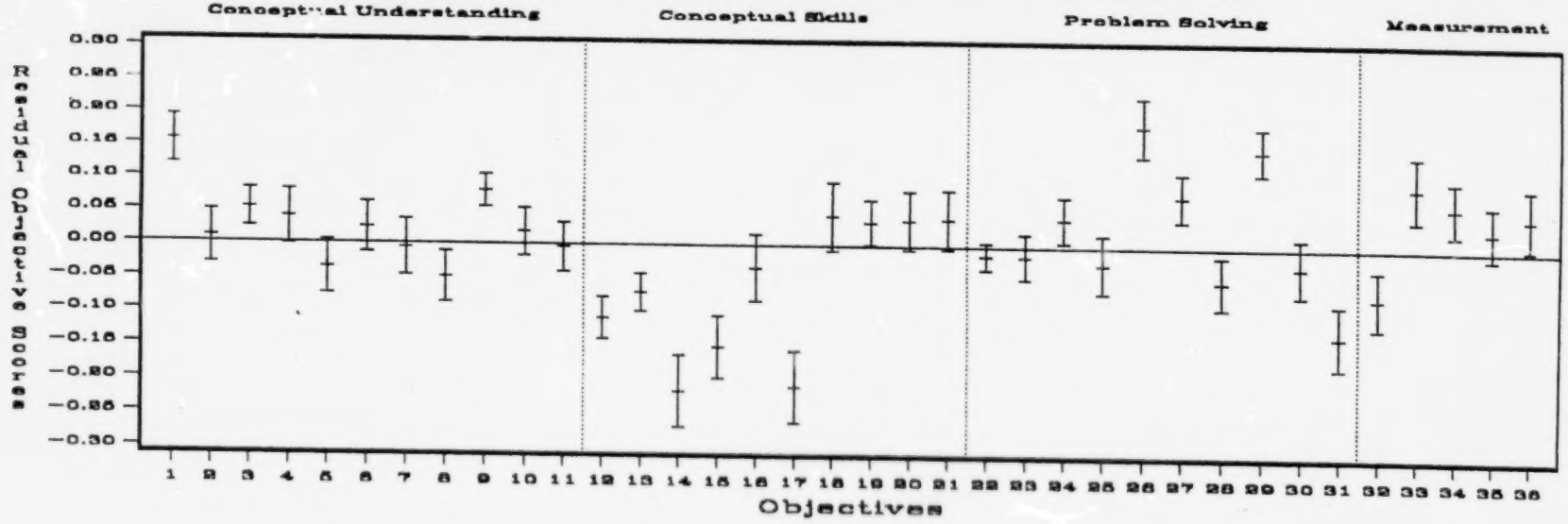


Figure 8. BLACK MALE RESIDUAL MEANS

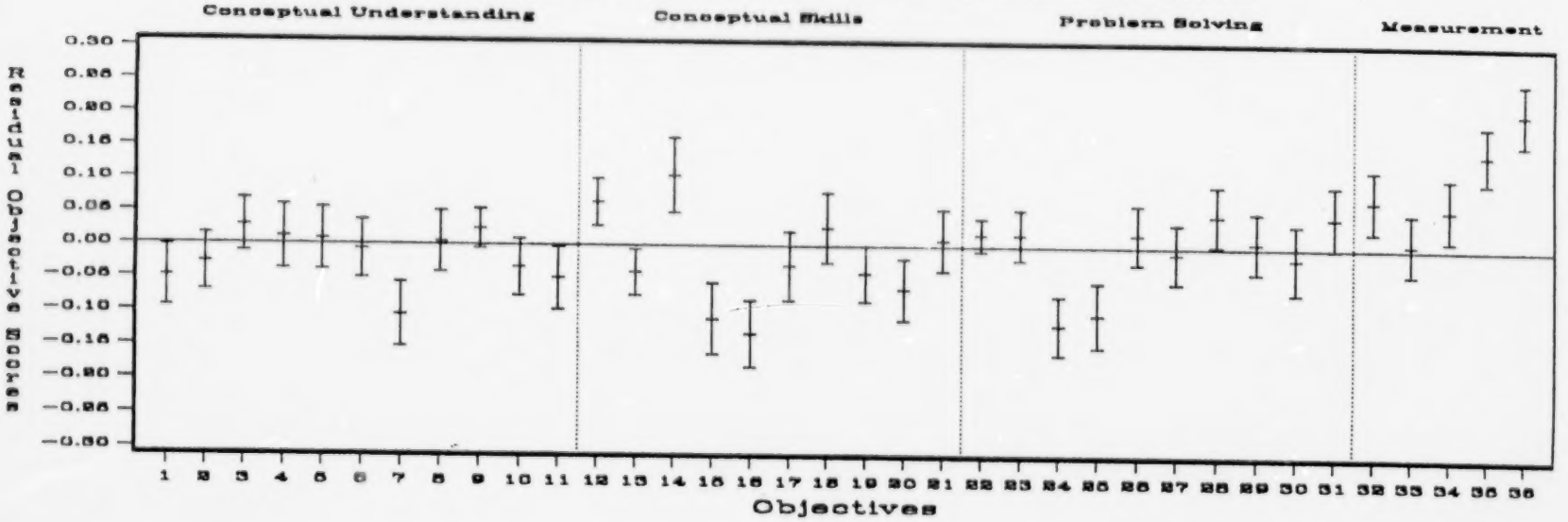


Figure 9. WHITE FEMALE RESIDUAL MEANS

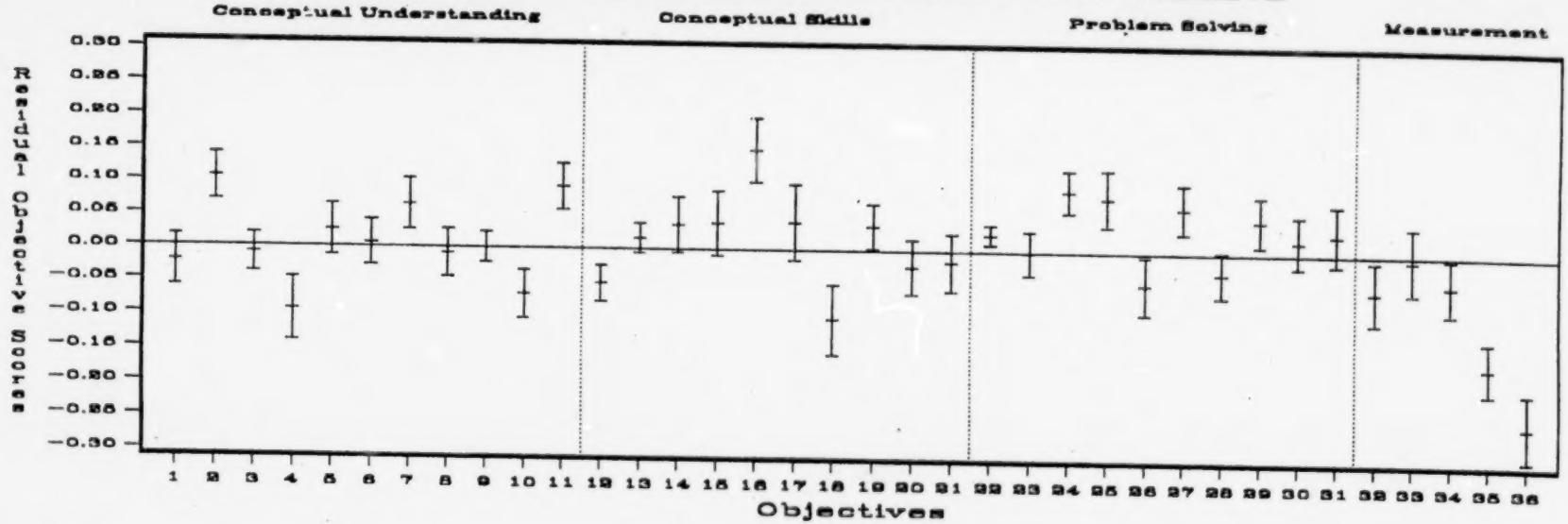
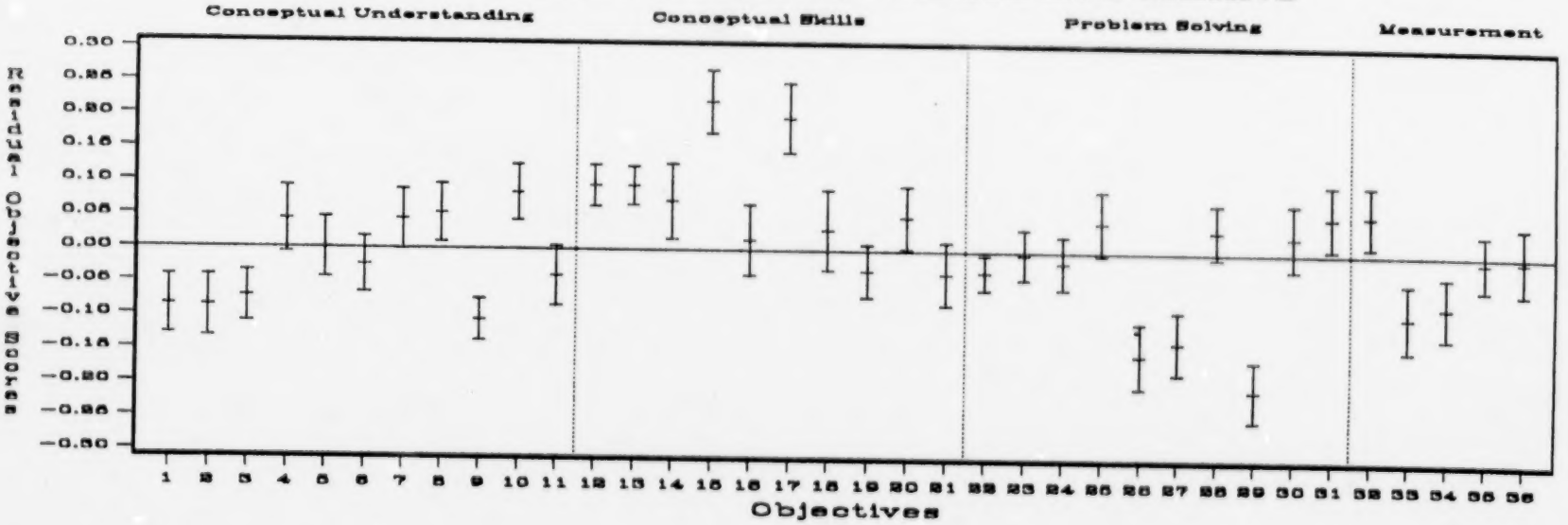


Figure 10. BLACK FEMALE RESIDUAL MEANS



Each performance profile displays the 36 residual mean objective scores for a particular demographic group. Objectives which tend to disfavor a group are characterized by negative residual mean objective scores. Conversely, objectives that favor a group are characterized by positive residual mean objective scores. In these performance profiles, objective bands were constructed with ±1 standard error around the residual mean objective score. Objective bands that included a zero residual mean objective score were classified as "0," neither favoring nor disfavoring a group. Objective bands located above the zero residual mean were classified as "+," favoring the demographic subgroup; objective bands that fell below the zero residual mean were classified as "-," disfavoring the group. These performance profiles display the relative strengths and weaknesses of a demographic group by content domains and for objectives which comprise each of the domains. These outcomes are summarized at the level of content domains in Table 2.

Table 2
Group Objective Performance Summaries by Content Domain

GROUP	CONTENT DOMAIN											
	CONCEPTUAL UNDERSTANDING			COMPUTATIONAL SKILLS			PROBLEM SOLVING/ APPLICATION			MEASUREMENT		
	+	-	0	+	-	0	+	-	0	+	-	0
White Male	3	1	7	1	5	4	4	2	4	3	1	1
Black Male	0	1	10	2	4	4	0	2	8	4	0	1
White Female	3	2	6	2	2	6	5	1	4	0	3	2
Black Female	2	4	5	5	0	5	0	3	7	1	2	2

\+ = number of objectives in Content Domain that favors the group
\- = number of objectives in Content Domain that disfavors the group
0 = number of objectives in Content Domain neither favors nor disfavors the group

Performance profiles can uncover content-based information that significance tests alone cannot. Objectives which fail to show statistically significant *dof* are not necessarily void of potential bias. An analysis of the group performance profiles shows, for example, that Objectives 13 and 15 both disfavored White Males and Black Males,

favored Black Females, and neither favored nor disfavored White Females. Although these outcomes were not statistically significant for *dof*, the objectives appeared to disfavor Males overall as a group. The performance summaries illustrate how *dof* can be used to diagnose performance at the content level. These methods and examples do not, however, diminish the necessity for a substantive review of the content.

SUMMARY AND DISCUSSION

The concept of *dif* was extended to the content unit. The interpretation of *dof* was illustrated with examples of statistically significant *dof* interactions in the context of item-level *dif* data. In the presence of significant *dof* and consistent patterns of interactions at the item- and the objective-levels, *dof* is attributable to content-based factors. Group performance profiles were constructed for each demographic subgroup in the study. These profiles identified the relative strengths and weaknesses of objective level performance by separate subgroups. Substantive information about potentially biased curricular objectives was detected between different group performance profiles. Content-based data can be used for diagnostic purposes; they can also augment item-level *dif* data and help contextualize statistical significance.

According to Bauer (1992), local test development activities continue at a high level. A critical step in test development is the identification of potentially biased items that favor one group of examinees independent of ability level. As discussed in Skaggs and Lissitz's study (1992) of consistency in item bias detection, *dif* can consistently flag items for no apparent reason. Differences in instructional background and opportunity to learn can be confounded with differences in matched-ability group performance. Based on collateral item information, *dof* can identify objectives that consistently yield aberrant results from expected performance at the objective level across different demographic groups.

Recent surveys of test use (Bauer, 1992; Nolen, Haladyna, & Haas, 1992) reported that the majority of local school districts and classroom teachers used tests for diagnostic and instructional purposes. If one of the primary purposes of testing is to provide

information about the success of instructional delivery or to identify curricular areas for remediation, test results should also provide diagnostic information to satisfy these goals: This diagnostic information must necessarily be content-based. As illustrated in this study, *dof* can be used to create group performance profiles by instructional units to target the relative strengths and weaknesses of demographic groups according to tested objectives.

One direction for future research is to explore the effect of multidimensionality on the sensitivity of *dof*. Test items are usually categorized into different content-based units with the assumption that each content-based unit is conceptually distinct.

Another methodological direction for future research is to explore a hierarchical structure for *dof* analysis. The test blueprint has an inherent structure of test items within content objectives, nested within content domains. The dependencies between and within nested units may be explicitly modeled through hierarchical methods.

Although data about opportunity to learn and differences in other curricular factors were not available for this study, inclusion of those types of data can only lead to more comprehensive and informed inferences about curricular outcomes.

REFERENCES

Bauer, E.A. (1992). NATD survey of testing practices and issues. *Educational Measurement: Issues and Practice, 11*(1), 10-14.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 147-200). New York: Macmillan.

Ironson, G.H. (1983). Using item response theory to measure bias. In R.K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Lehman, J.D. (1986). Opportunity to learn and differential item functioning. Unpublished doctoral dissertation, University of California, Los Angeles.

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109-118.

Muthén, B.O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28(1),* 1-22

Nolen, S.B., Haladyna, T.M., & Haas, N.S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice, 11*(2), 9-15.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317-375.

Skaggs, G. & Lissitz, R.W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement, 29*(3), 227-242.

Tang, H. (1994, April). *A simultaneous approach to multi-factor DIF analysis.* Paper presented at the 1994 Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Wright, B.D., Mead, R., & Draba, R. (1976). Detecting and correcting item bias with a logistic response model. Research Memorandum No. 22. Chicago: University of Chicago.

APPENDIX A

Dof Statistics by Demographic Subgroups

DOF DATA, 2-WAY INTERACTIONS
n = 400

OBJ	DOMAIN	NAME	MEAN RAW OBJECTIVE SCORES Black Female	Black Male	White Female	White Male	RESIDUAL MEAN OBJECTIVE SCORES Black Female	Black Male	White Female	White Male	RES MEAN MAGNITUDE Gender Gap (B vs. W)	Ethnic Gap (M vs. F)	F	P
1	C	ORDER FRACTIONS	2.39	2.37	3.01	3.16	-0.09	-0.05	-0.02	0.16	0.14	0.14	2.90	0.089
2	C	ORDER DECIMALS	2.71	2.71	3.39	3.26	-0.09	-0.03	0.11	0.01	0.04	0.16	3.70	0.056
3	C	ROUND WHOLE NUMBER	3.29	3.33	3.66	3.69	-0.07	0.03	-0.01	0.05	0.04	0.04	0.30	0.562
4	C	RND DECIMALS TO NE	2.66	2.55	3.11	3.20	0.04	0.01	-0.09	0.04	0.10	0.10	3.10	0.078
5	C	MUL/DIV WHL #S & D	2.58	2.53	3.15	3.06	0.00	0.01	0.03	-0.04	0.06	0.02	0.60	0.421
6	C	ID FRACT, DEC AND	2.34	2.30	2.98	2.96	-0.02	-0.01	0.01	0.02	0.00	0.00	0.00	0.995
7	C	CONVERT FRACT TO D	2.85	2.63	3.39	3.28	0.04	-0.10	0.07	-0.01	0.06	0.06	0.80	0.367
8	C	CNVT FRACTS & DEC	2.72	2.61	3.22	3.14	0.05	0.00	-0.01	-0.05	0.01	0.01	0.00	0.921
9	C	ID PTS ON NUM LINE	3.31	3.41	3.67	3.73	-0.11	0.02	0.00	0.08	0.05	0.05	0.90	0.351
10	C	ID RATIOS AND FRAC	2.84	2.67	3.17	3.23	0.08	-0.03	-0.07	0.02	0.02	0.10	6.90	0.009
11	C	ID EST PROC WITH F	2.65	2.57	3.33	3.20	-0.04	-0.05	0.09	0.00	0.08	0.08	1.10	0.288
12	P	ADD/SUBT WHLE #S L	3.65	3.57	3.73	3.64	0.09	0.07	-0.05	-0.11	0.04	0.04	0.20	0.650
13	P	MULT/DIV 2,3 DIGT	3.71	3.53	3.83	3.71	0.10	-0.04	0.02	-0.07	0.05	0.05	0.70	0.403
14	P	ADD/SUBT DEC TO .X	3.30	3.27	3.64	3.34	0.07	0.11	0.04	-0.22	0.22	0.30	7.70	0.006
15	P	ID CORR PLACE OF D	2.59	2.19	3.04	2.82	0.22	-0.11	0.04	-0.15	0.14	0.14	2.10	0.152
16	P	ADD/SUBT FRACTS AN	1.84	1.65	2.69	2.49	0.01	-0.13	0.15	-0.03	0.04	0.04	0.10	0.709
17	P	MULTIPLY FRACTS AN	1.95	1.68	2.50	2.24	0.20	-0.03	0.04	-0.21	0.02	0.02	0.00	0.824
18	P	DETERMINE THE % OF	1.80	1.75	2.41	2.54	0.03	0.03	-0.10	0.04	0.14	0.12	1.90	0.164
19	P	EST SUM/DIFF OF WH	2.92	2.84	3.45	3.41	-0.03	-0.04	0.04	0.04	0.01	0.01	0.00	0.898
20	P	EST PROD/QUOT OF W	2.27	2.10	2.84	2.87	0.05	-0.07	-0.02	0.04	0.06	0.04	4.00	0.046
21	P	EST FRACT PTS & %	2.00	1.99	2.72	2.75	-0.03	0.01	-0.02	0.04	0.02	0.02	0.00	0.882
22	A	COMP SUMS/DIFF/PRO	3.72	3.74	3.91	3.85	-0.03	0.02	0.03	-0.01	0.01	0.03	4.00	0.046
23	A	INTERPRET GRAPHS,	2.92	2.89	3.34	3.30	0.00	0.02	0.00	-0.01	0.01	0.03	0.20	0.626
24	A	SOLVE 1-2 STP PROB	2.89	2.72	3.47	3.38	-0.02	-0.12	0.09	0.04	0.05	0.05	0.40	0.504
25	A	SOLVE 1-2 STP PROB	2.15	1.95	2.89	2.75	0.04	-0.10	0.08	-0.02	0.04	0.04	0.20	0.675
26	A	SOLVE PROBS INVOLV	1.28	1.43	2.10	2.33	-0.15	0.02	-0.05	0.18	0.06	0.06	0.40	0.527
27	A	SOLVE PROBS INVOLV	2.31	2.39	3.05	3.04	-0.13	-0.01	0.07	0.08	0.11	0.11	2.00	0.153
28	A	EST REASONABLE ANS	2.78	2.73	3.24	3.18	0.03	0.05	-0.03	-0.05	0.00	0.04	0.20	0.646
29	A	SOLVE EXTRANEOUS I	2.29	2.43	3.14	3.20	-0.21	0.01	0.05	0.15	0.12	0.12	2.00	0.154
30	A	ID NEEDED INFO IN	2.62	2.51	3.21	3.12	0.02	-0.02	0.02	-0.03	0.01	0.01	0.00	0.947
31	A	SOLVE PROCESS PROB	2.04	1.98	2.72	2.53	0.05	0.05	0.03	-0.13	0.16	0.16	2.70	0.101
32	M	ID FIGURES USING G	2.20	2.17	2.74	2.69	0.06	0.07	-0.06	-0.07	0.00	0.02	0.10	0.706
33	M	MEASURE/DETERMINE	1.46	1.53	2.28	2.37	-0.09	0.01	-0.01	0.09	0.00	0.00	0.00	0.976
34	M	EST LENGTHS/AREAS/	2.33	2.41	2.98	3.05	-0.08	0.06	-0.04	0.06	0.04	0.04	0.10	0.713
35	M	PICK APPROP METRIC	2.79	2.89	3.11	3.27	-0.01	0.15	-0.16	0.03	0.03	0.03	0.20	0.637
36	M	MAKE MEAS CONVERSI	1.74	1.91	2.25	2.53	-0.01	0.21	-0.25	0.05	0.08	0.08	0.80	0.372

DOF DATA, ETHNICITY MAIN EFFECTS
n = 800

			MEAN RAW OBJECTIVE SCORES		RESIDUAL MEAN OBJECTIVE SCORES				
OBJ	DOMAIN	NAME	Black	White	Black	White	F	P	signif
1	C	ORDER FRACTIONS	2.38	3.08	-0.07	0.07	10.60	0.001	*
2	C	ORDER DECIMALS	2.71	3.32	-0.06	0.06	8.00	0.005	*
3	C	ROUND WHOLE NUMBER	3.31	3.67	-0.02	0.02	1.70	0.197	
4	C	RND DECIMALS TO NE	2.61	3.15	0.03	-0.03	1.30	0.248	
5	C	MUL/DIV WHL #S & D	2.55	3.10	0.00	0.00	0.00	0.824	
6	C	ID FRACT, DEC AND	2.32	2.97	-0.02	0.02	0.60	0.428	
7	C	CONVERT FRACT TO D	2.74	3.33	-0.03	0.03	2.00	0.160	
8	C	CNVT FRACTS & DEC	2.66	3.18	0.03	-0.03	1.90	0.164	
9	C	ID PTS ON NUM LINE	3.36	3.70	-0.04	0.04	9.10	0.003	*
10	C	ID RATIOS AND FRAC	2.75	3.20	0.03	-0.03	1.70	0.196	
11	C	ID EST PROC WITH F	2.61	3.26	-0.04	0.04	4.80	0.029	*
12	P	ADD/SUBT WHLE #S L	3.61	3.68	0.08	-0.08	26.40	0.000	*
13	P	MULT/DIV 2,3 DIGT	3.62	3.77	0.03	-0.03	3.50	0.060	
14	P	ADD/SUBT DEC TO .X	3.29	3.49	0.09	-0.09	11.90	0.001	#
15	P	ID CORR PLACE OF D	2.39	2.93	0.06	-0.06	5.20	0.022	*
16	P	ADD/SUBT FRACTS AN	1.74	2.59	-0.06	0.06	5.50	0.019	*
17	P	MULTIPLY FRACTS AN	1.81	2.37	0.08	-0.08	9.60	0.002	*
18	P	DETERMINE THE % OF	1.77	2.47	0.03	-0.03	1.20	0.279	
19	P	EST SUM/DIFF OF WH	2.88	3.43	-0.04	0.04	3.80	0.051	
20	P	EST PROD/QUOT OF W	2.19	2.85	-0.01	0.01	0.10	0.730	
21	P	EST FRACT PTS & %	1.99	2.74	-0.01	0.01	0.30	0.583	
22	A	COMP SUMS/DIFF/PRO	3.73	3.88	-0.01	0.01	0.40	0.546	
23	A	INTERPRET GRAPHS,	2.91	3.32	0.01	-0.01	0.20	0.686	
24	A	SOLVE 1-2 STP PROB	2.80	3.42	-0.07	0.07	12.80	0.000	*
25	A	SOLVE 1-2 STP PROB	2.05	2.82	-0.03	0.03	1.60	0.200	
26	A	SOLVE PROBS INVOLV	1.36	2.22	-0.07	0.07	8.80	0.003	*
27	A	SOLVE PROBS INVOLV	2.35	3.04	-0.07	0.07	12.40	0.000	*
28	A	EST REASONABLE ANS	2.75	3.21	0.04	-0.04	4.20	0.042	*
29	A	SOLVE EXTRANEOUS I	2.36	3.17	-0.10	0.10	23.20	0.000	*
30	A	ID NEEDED INFO IN	2.57	3.16	0.00	0.00	0.00	0.858	
31	A	SOLVE PROCESS PROB	2.01	2.62	0.05	-0.05	4.70	0.031	*
32	M	ID FIGURES USING G	2.19	2.71	0.06	-0.06	8.00	0.005	*
33	M	MEASURE/DETERMINE	1.50	2.32	-0.04	0.04	3.10	0.078	
34	M	EST LENGTHS/AREAS/	2.37	3.02	-0.01	0.01	0.20	0.682	
35	M	PICK APPROP METRIC	2.84	3.19	0.07	-0.07	11.30	0.001	*
36	M	MAKE MEAS CONVERSI	1.83	2.39	0.10	-0.10	17.90	0.000	*

= Presence of significant two-way interaction effect

DOF DATA, GENDER MAIN EFFECTS
n = 800

			MEAN RAW OBJECTIVE SCORES		RESIDUAL MEAN OBJECTIVE SCORES				
OBJ	DOMAIN	NAME	Female	Male	Female	Male	F	P	signif
1	C	ORDER FRACTIONS	2.70	2.76	-0.05	0.05	6.70	0.010	*
2	C	ORDER DECIMALS	3.05	2.98	0.01	-0.01	0.20	0.654	
3	C	ROUND WHOLE NUMBER	3.47	3.51	-0.04	0.04	5.70	0.017	*
4	C	RND DECIMALS TO NE	2.88	2.87	-0.03	0.02	1.20	0.283	
5	C	MUL/DIV WHL #S & D	2.86	2.79	0.01	-0.01	0.40	0.525	
6	C	ID FRACT, DEC AND	2.66	2.63	-0.01	0.01	0.20	0.659	
7	C	CONVERT FRACT TO D	3.12	2.95	0.05	-0.05	6.40	0.011	*
8	C	CNVT FRACTS & DEC	2.97	2.87	0.02	-0.02	1.20	0.278	
9	C	ID PTS ON NUM LINE	3.49	3.57	-0.05	0.05	15.10	0.000	*
10	C	ID RATIOS AND FRAC	3.00	2.95	0.01	-0.01	0.20	0.694	
11	C	ID EST PROC WITH F	2.99	2.88	0.03	-0.03	1.70	0.195	
12	P	ADD/SUBT WHLE #S L	3.69	3.61	0.02	-0.02	1.90	0.167	
13	P	MULT,/DIV 2,3 DIGT	3.77	3.62	0.06	-0.06	15.10	0.000	*
14	P	ADD/SUBT DEC TO .X	3.47	3.31	0.05	-0.06	4.50	0.035	#
15	P	ID CORR PLACE OF D	2.82	2.50	0.13	-0.13	28.70	0.000	*
16	P	ADD/SUBT FRACTS AN	2.27	2.07	0.08	-0.08	10.90	0.001	*
17	P	MULTIPLY FRACTS AN	2.22	1.96	0.12	-0.12	20.00	0.000	*
18	P	DETERMINE THE % OF	2.10	2.14	-0.04	0.04	1.80	0.180	
19	P	EST SUM/DIFF OF WH	3.18	3.12	0.00	0.00	0.00	0.860	
20	P	EST PROD/QUOT OF W	2.55	2.49	0.01	-0.01	0.40	0.550	
21	P	EST FRACT PTS & %	2.36	2.37	-0.03	0.02	1.20	0.269	
22	A	COMP SUMS/DIFF/PRO	3.82	3.80	0.00	0.00	0.00	0.825	
23	A	INTERPRET GRAPHS,	3.13	3.10	0.00	0.00	0.00	0.897	
24	A	SOLVE 1-2 STP PROB	3.18	3.05	0.04	-0.04	4.20	0.042	*
25	A	SOLVE 1-2 STP PROB	2.52	2.35	0.06	-0.06	7.80	0.005	*
26	A	SOLVE PROBS INVOLV	1.69	1.88	-0.10	0.10	19.70	0.000	*
27	A	SOLVE PROBS INVOLV	2.68	2.71	-0.03	0.03	2.60	0.105	
28	A	EST REASONABLE ANS	3.01	2.95	0.00	0.00	0.00	0.959	
29	A	SOLVE EXTRANEOUS I	2.71	2.82	-0.08	0.08	14.30	0.000	*
30	A	ID NEEDED INFO IN	2.92	2.81	0.02	-0.02	0.90	0.333	
31	A	SOLVE PROCESS PROB	2.38	2.26	0.04	-0.04	3.30	0.070	
32	M	ID FIGURES USING G	2.47	2.43	0.00	0.00	0.00	0.969	
33	M	MEASURE/DETERMINE	1.87	1.95	-0.05	0.05	4.20	0.041	*
34	M	EST LENGTHS/AREAS/	2.66	2.73	-0.06	0.06	8.00	0.005	*
35	M	PICK APPROP METRIC	2.95	3.08	-0.09	0.09	18.50	0.000	*
36	M	MAKE MEAS CONVERSI	1.99	2.22	-0.13	0.13	28.80	0.000	*

= Presence of significant two-way interaction effect

APPENDIX B

Item-Level *dif* Statistics for *dof* Objectives #3, 10, 14

ITEM-LEVEL DIF DATA, 2-WAY INTERACTIONS
OBJECTIVE #3: Round Whole Number
n = 400

	MEAN RAW ITEM SCORES				RESIDUAL MEAN ITEM SCORES				STANDARD ERROR RESIDUAL MEAN ITEM SCORES						
ITEM	Black Female	Black Male	White Female	White Male	Black Female	Black Male	White Female	White Male	Black Female	Black Male	White Female	White Male	F	P	signif
105	0.89	0.91	0.96	0.95	-0.02	0.01	0.00	0.00	0.02	0.01	0.01	0.01	1.6	0.209	
106	0.90	0.86	0.94	0.95	0.01	-0.02	0.00	0.01	0.01	0.02	0.01	0.01	2.1	0.151	
107	0.84	0.86	0.94	0.94	-0.02	0.01	0.00	0.01	0.02	0.02	0.01	0.01	0.4	0.530	
108	0.66	0.71	0.82	0.85	-0.04	0.03	-0.01	0.03	0.02	0.02	0.02	0.02	0.5	0.471	

OBJECTIVE #10: Identify Ratios and Fractions from Pictures
n = 400

	MEAN RAW ITEM SCORES				RESIDUAL MEAN ITEM SCORES				STANDARD ERROR RESIDUAL MEAN ITEM SCORES						
ITEM	Black Female	Black Male	White Female	White Male	Black Female	Black Male	White Female	White Male	Black Female	Black Male	White Female	White Male	F	P	signif
113	0.93	0.88	0.96	0.96	0.02	-0.02	0.00	0.01	0.01	0.02	0.01	0.01	3.6	0.056	
114	0.86	0.82	0.91	0.91	0.02	-0.01	-0.01	0.00	0.02	0.02	0.01	0.01	0.8	0.364	
115	0.63	0.55	0.70	0.67	0.07	0.01	-0.03	-0.05	0.02	0.02	0.02	0.02	0.8	0.379	
116	0.43	0.42	0.61	0.69	-0.02	-0.02	-0.03	0.06	0.02	0.02	0.02	0.02	4.6	0.031	*

OBJECTIVE #14: Add/Subtract Decimals to .XX -- Horizontally
n = 400

	MEAN RAW ITEM SCORES				RESIDUAL MEAN ITEM SCORES				STANDARD ERROR RESIDUAL MEAN ITEM SCORES						
ITEM	Black Female	Black Male	White Female	White Male	Black Female	Black Male	White Female	White Male	Black Female	Black Male	White Female	White Male	F	P	signif
65	0.78	0.78	0.89	0.81	0.02	0.03	0.01	-0.06	0.02	0.02	0.01	0.02	4.9	0.028	*
66	0.83	0.83	0.91	0.82	0.03	0.03	0.01	-0.07	0.02	0.02	0.01	0.02	6.2	0.013	*
67	0.86	0.85	0.94	0.87	0.01	0.01	0.01	-0.04	0.02	0.02	0.01	0.02	3.3	0.069	
68	0.83	0.83	0.91	0.84	0.02	0.03	0.00	-0.05	0.02	0.02	0.01	0.02	4.2	0.041	*